nature portfolio

Corresponding author(s): Xiaochen Bo

Last updated by author(s): Jun 29, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The source data and codes of EviDTI are available on GitHub at https://github.com/zhaoyanpeng208/EviDTI, which has also been deposited in the Zenodo under accession code https://zenodo.org/records/15760471 This implementation is based on the python 3.8.20 environment. numpy >= 1.24.4 pandas >= 1.4.1
Data analysis	The source data and codes of EviDTI are available on GitHub at https://github.com/zhaoyanpeng208/EviDTI, which has also been deposited in the Zenodo under accession code https://zenodo.org/records/15760471 Prottrans (https://github.com/agemagician/ProtTrans) is used for calculating target protein representations. Data are analyzed using numpy v1.24.3 (https://numpy.org/), pandas v1.4.2 (https://pandas.pydata.org/), Seaborn V0.11.2 (https://seaborn.pydata.org/) and Matplotlib v3.4.3 (https://matplotlib.org/). Structures are visualized by Pymol v2.5.2 (https://www.pymol.org/).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The source data of three datasets used to train and evaluate the model is provided at https://github.com/zhaoyanpeng208/EviDTI/tree/main/dataset and https://zenodo.org/records/14056305. The source data of 67 targets with 51 drugs used for literature validation of tyrosine kinase modulators is provided in https://github.com/zhaoyanpeng208/EviDTI/tree/main/dataset/len_v_s_ dataset and https://zenodo.org/records/14056305. The source data of two Lenvatinib analogues (LYD-2-45 and LYD-2-49) with 11 tyrosine kinase targets used for patent validation of tyrosine kinase modulators is provided in https://github.com/zhaoyanpeng208/EviDTI/tree/main/dataset/len_case_ dataset and https://zenodo.org/records/14056305. Source data are provided with this paper through Figshare https://doi.org/10.6084/m9.figshare.28816634.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	To evaluate the effectiveness of the EviDTI framework, we validated it on three different experimental datasets: DrugBank, Davis, and KIBA. These three datasets are commonly used benchmark datasets for DTI prediction. DrugBank dataset: A total of 33,062 pairs of samples were collected in the DrugBank dataset, with a ratio of positive to negative samples of 1:1. Davis dataset: A total of 25,772 pairs of samples were used for the model building. The data sets contained 7320 positive pairs and 18452 negative pairs. KIBA dataset: A total of 116,350 pairs of samples were used for the model building. The data sets contained 22,154 positive pairs and 94,196 negative pairs.
Data exclusions	Drugs with inorganic compounds, very small molecule compounds [e.g., iron (DB01592), zinc (DB01593)], or SMILES strings that the RDKit Python package could not recognize were removed. Drugs with a SMILES length greater than 300 were also removed to comply with the maximum input of MG-BERT.
Replication	We have provided detailed instructions in our codes to reproduce experiment results.
Randomization	All datasets are randomly divided into training, validation and test sets in the ratio of 8:1:1
Blinding	The investigators were blinded to group allocation during experiments and outcome analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A